L10: Probability, statistics, and estimation theory

Review of probability theory

Bayes theorem

Statistics and the Normal distribution

Least Squares Error estimation

Maximum Likelihood estimation

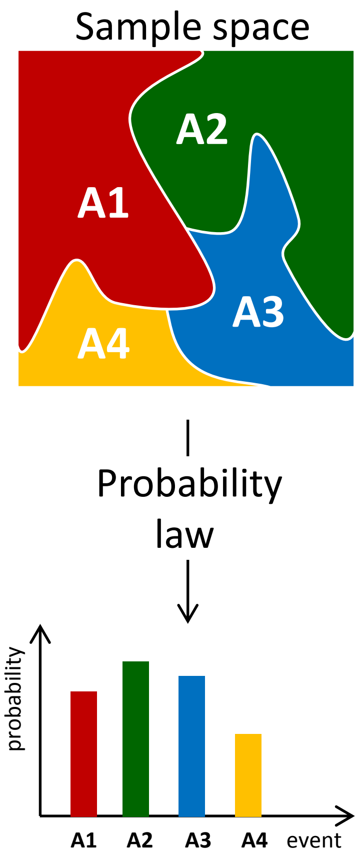
Bayesian estimation

This lecture is partly based on [Huang, Acero and Hon, 2001, ch. 3]

Review of probability theory

Definitions (informal)

- Probabilities are numbers assigned to events that indicate "***how likely***" it is that the event will occur when a random experiment is performed
- A probability law for a random experiment is a rule that assigns probabilities to the events in the experiment
- The sample space S of a random experiment is the set of all possible outcomes

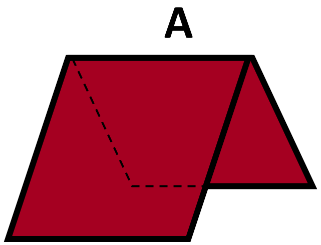


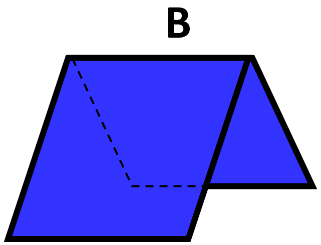
Axioms of probability

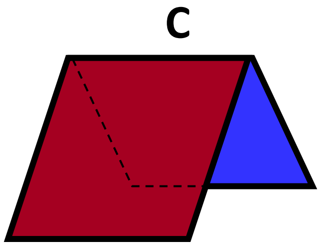
- Axiom I: $P[A_i] \geq 0$
- Axiom II: $P[S] = 1$
- Axiom III: $A_i \cap A_j = \emptyset \Rightarrow P[A_i \cup A_j] = P[A_i] + P[A_j]$

Warm-up exercise

- I show you three colored cards
  - One BLUE on both sides
  - One RED on both sides
  - One BLUE on one side, RED on the other







- I shuffle the three cards, then pick one and show you one side only. The side visible to you is RED
  - Obviously, the card has to be either A or C, *right*?
- I am willing to bet $1 that the other side of the card has the same color, and need someone in class to bet another $1 that it is the other color
  - On the average we will end up even, *right*?
  - Let’s try it!

More properties of probability

- $P[A^C] = 1 - P[A]$
- $P[A] \leq 1$
- $P[\emptyset] = 0$
- $given\ \{A_1 \dots A_N\}, \{A_i \cap A_j = \emptyset, \forall ij\} \Rightarrow P\left[\bigcup_{k=1}^{N} A_k\right] = \sum_{k=1}^{N} P[A_k]$
- $P[A_1 \cup A_2] = P[A_1] + P[A_2] - P[A_1 \cap A_2]$
- $P\left[\bigcup_{k=1}^{N} A_k\right] =$
  $\sum_{k=1}^{N} P[A_k] - \sum_{j<k}^{N} P\left[A_j \cap A_k\right] + \cdots + (-1)^{N+1} P[A_1 \cap A_2 \dots \cap A_N]$
- $A_1 \subset A_2 \Rightarrow P[A_1] \leq P[A_2]$

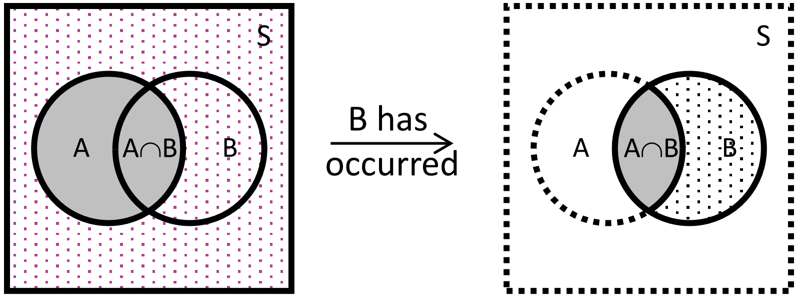
Conditional probability

- If A and B are two events, the probability of event A when we already know that event B has occurred is

$$P[A|B] = \frac{P[A \cap B]}{P[B]} \quad if \;\; P[B] > 0$$

  - This conditional probability P[A|B] is read:
    - the “conditional probability of A conditioned on B”, or simply
    - the “probability of A given B”

- Interpretation
  - The new evidence “*B has occurred*” has the following effects
    - The original sample space S (the square) becomes B (the rightmost circle)
    - The event A becomes A∩B
  - P[B] simply re-normalizes the probability of events that occur jointly with B



Theorem of total probability

- Let $B_1, B_2 \ldots B_N$ be a partition of S (mutually exclusive that add to S)
- Any event A can be represented as

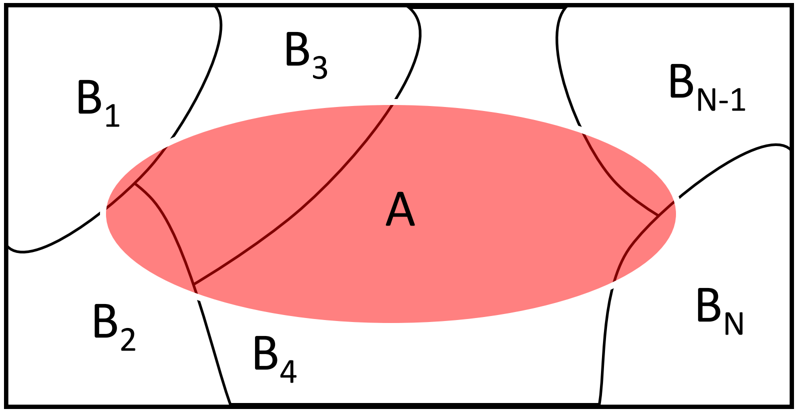
$$A = A \cap S = A \cap (B_1 \cup B_2 \ldots B_N) = (A \cap B_1) \cup (A \cap B_2) \ldots (A \cap B_N)$$

- Since $B_1, B_2 \ldots B_N$ are mutually exclusive, then

$$P[A] = P\{A \cap B_1] + P\{A \cap B_2] + \cdots + P\{A \cap B_N]$$

- and, therefore

$$P[A] = P[A|B_1]P[B_1] + \cdots P[A|B_N]P[B_N] = \sum_{k=1}^{N} P[A|B_k]P[B_k]$$



Bayes theorem

- Assume $\{B_1, B_2 \ldots B_N\}$ is a partition of S
- Suppose that event A occurs
- What is the probability of event B_j?

- Using the definition of conditional probability and the Theorem of total probability we obtain

$$P[B_j|A] = \frac{P[A \cap B_j]}{P[A]} = \frac{P[A|B_j]P[B_j]}{\sum_{k=1}^{N} P[A|B_k]P[B_k]}$$

- This is known as Bayes Theorem or Bayes Rule, and is (one of) the most useful relations in probability and statistics

Bayes theorem and statistical pattern recognition

– When used for pattern classification, BT is generally expressed as

$$P[\omega_j|x] = \frac{P[x|\omega_j]P[\omega_j]}{\sum_{k=1}^{N} P[x|\omega_k]P[\omega_k]} = \frac{P[x|\omega_j]P[\omega_j]}{P[x]}$$

- where ω_j is the j-th class (e.g., phoneme) and x is the feature/observation vector (e.g., vector of MFCCs)

– A typical decision rule is to choose class ω_j with highest $\mathrm{P}[\omega_j|x]$

- Intuitively, we choose the class that is more “likely” given observation x

– Each term in the Bayes Theorem has a special name

- $P[\omega_j]$ prior probability (of class ω_j)
- $P[\omega_j|x]$ posterior probability (of class ω_j given the observation x)
- $P[x|\omega_j]$ likelihood (probability of observation x given class ω_j)
- $P[x]$ normalization constant (does not affect the decision)

Example

- Consider a clinical problem where we need to decide if a patient has a particular medical condition on the basis of an imperfect test
  - Someone with the condition may go undetected (false-negative)
  - Someone free of the condition may yield a positive result (false-positive)
- Nomenclature
  - The true-negative rate P(NEG|¬COND) of a test is called its SPECIFICITY
  - The true-positive rate P(POS|COND) of a test is called its SENSITIVITY
- Problem
  - Assume a population of 10,000 with a 1% prevalence for the condition
  - Assume that we design a test with 98% specificity and 90% sensitivity
  - Assume you take the test, and the result comes out POSITIVE
  - What is the probability that you have the condition?
- Solution
  - Fill in the joint frequency table next slide, or
  - Apply Bayes rule

	TEST IS POSITIVE	**TEST IS NEGATIVE**	**ROW TOTAL**
HAS CONDITION	*True-positive* *P(POS|COND)*	*False-negative* *P(NEG|COND)*	
FREE OF CONDITION	*False-positive* *P(POS|¬COND)*	*True-negative* *P(NEG|¬COND)*	
COLUMN TOTAL			

	TEST IS POSITIVE	TEST IS NEGATIVE	ROW TOTAL
HAS CONDITION	*True-positive* *P(POS|COND)* 100×0.90	*False-negative* *P(NEG|COND)* 100×(1-0.90)	100
FREE OF CONDITION	*False-positive* *P(POS|¬COND)* 9,900×(1-0.98)	*True-negative* *P(NEG|¬COND)* 9,900×0.98	9,900
COLUMN TOTAL	288	9,712	10,000

- Applying Bayes rule

$$P[cond|+] =$$

$$= \frac{P[+|cond]P[cond]}{P[+]} =$$

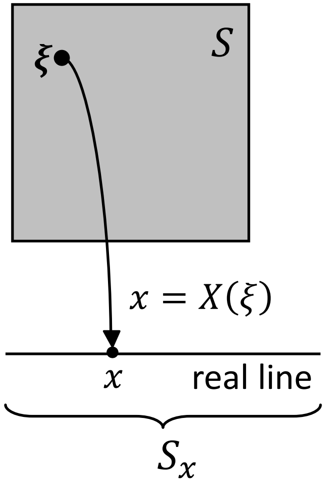
$$= \frac{P[+|cond]P[cond]}{P[+|cond]P[cond] + P[+|\neg cond]P[\neg cond]} =$$

$$= \frac{0.90 \times 0.01}{0.90 \times 0.01 + (1 - 0.98) \times 0.99} =$$

$$= 0.3125$$

Random variables

- When we perform a random experiment we are usually interested in some measurement or numerical attribute of the outcome
  - e.g., weights in a population of subjects, execution times when benchmarking CPUs, shape parameters when performing ATR
- These examples lead to the concept of random variable
  - A random variable X is a function that assigns a real number $X(\xi)$ to each outcome ξ in the sample space of a random experiment
  - $X(\xi)$ maps from all possible outcomes in sample space onto the real line
- The function that assigns values to each outcome is fixed and deterministic, i.e., as in the rule "count the number of heads in three coin tosses"
  - Randomness in X is due to the underlying randomness of the outcome ξ of the experiment
- Random variables can be
  - Discrete, e.g., the resulting number after rolling a dice
  - Continuous, e.g., the weight of a sampled individual

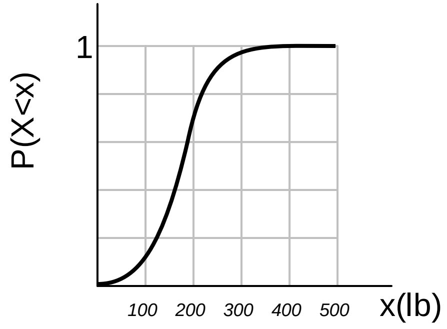


Cumulative distribution function (cdf)

- The cumulative distribution function $F_X(x)$ of a random variable X is defined as the probability of the event $\{X \le x\}$

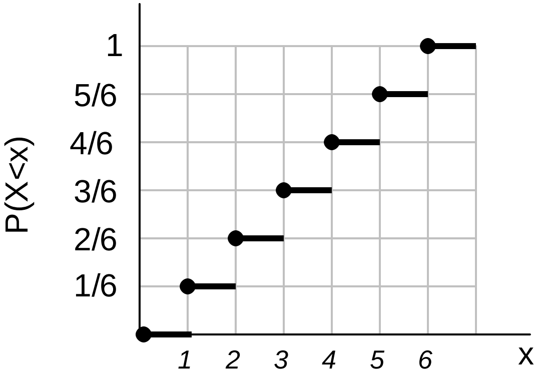
$$F_X(x) = P[X \le x\ ] \quad -\infty < x < \infty$$

- Intuitively, $F_X(b)$ is the long-term proportion of times when $X(\xi) \le b$



cdf for a person's weight

- Properties of the cdf
  - $0 \le F_X(x) \le 1$
  - $\lim\limits_{x\to\infty} F_X(x) = 1$
  - $\lim\limits_{x\to-\infty} F_X(x) = 0$
  - $F_X(a) \le F_X(b)\ if\ a \le b$
  - $\mathrm{F_X}(b) = \lim\limits_{h\to 0} F_X(b+h) = F_X(b^+)$

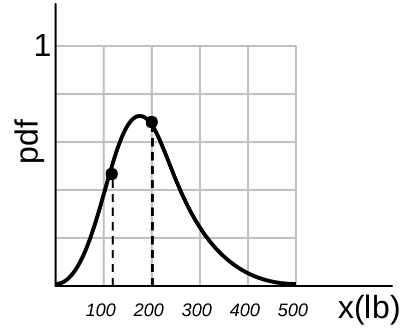


cdf for rolling a dice

Probability density function (pdf)

- The probability density function $f_X(x)$ of a continuous random variable X, if it exists, is defined as the derivative of $F_X(x)$

$$f_X(x) = \frac{dF_X(x)}{dx}$$

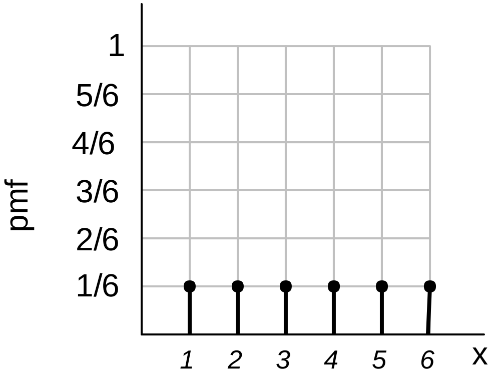


pdf for a person's weight

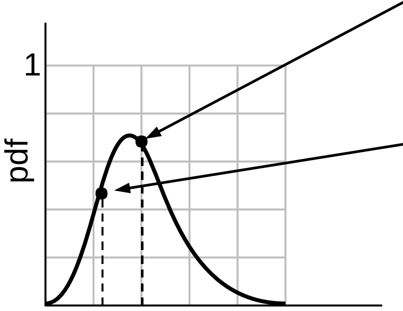
- For discrete random variables, the equivalent to the pdf is the probability mass function

$$f_X(x) = \frac{\Delta F_X(x)}{\Delta x}$$

- Properties
  - $f_X(x) > 0$
  - $P[a < x < b] = \int_a^b f_X(x)dx$
  - $F_X(x) = \int_{-\infty}^x f_X(x)dx$
  - $1 = \int_{-\infty}^{\infty} f_X(x)dx$
  - $f_X(x|A) = \frac{d}{dx}F_X(x|A) \quad where \quad F_X(x|A) = \frac{P[\{X<x\}\cap A]}{P[A]} \quad if \quad P[A] > 0$

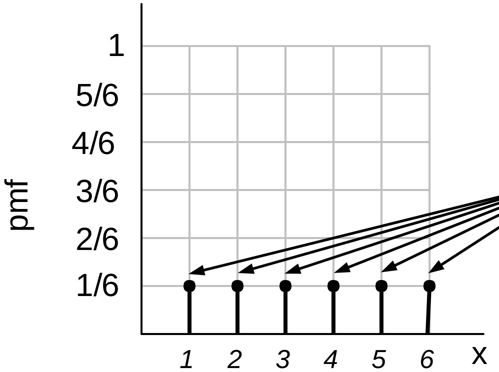


pmf for rolling a (fair) dice



pdf for a person's weight

- **What is the probability of somebody weighting 200 lb?**
  - According to the pdf, this is about 0.62
  - This number seems reasonable, right?
- **Now, what is the probability of somebody weighting 124.876 lb?**
  - According to the pdf, this is about 0.43
  - But, intuitively, we know that the probability should be zero (or very, very small)
- **How do we explain this paradox?**
  - The pdf DOES NOT define a probability, but a probability DENSITY!
  - To obtain the actual probability we must integrate the pdf in an interval
  - So we should have asked the question: what is the probability of somebody weighting 124.876 lb plus or minus 2 lb?



pmf for rolling a (fair) dice

- **The probability mass function is a 'true' probability (reason why we call it a 'mass' as opposed to a 'density')**
  - The pmf is indicating that the probability of any number when rolling a fair dice is the same for all numbers, and equal to 1/6, a very legitimate answer
  - The pmf DOES NOT need to be integrated to obtain the probability (it cannot be integrated in the first place)

Statistical characterization of random variables

- The cdf or the pdf are SUFFICIENT to fully characterize a r.v.
- However, a r.v. can be PARTIALLY characterized with other measures
- Expectation (center of mass of a density)

$$E[X] = \mu = \int_{-\infty}^{\infty} x f_X(x) dx$$

- Variance (spread about the mean)

$$var[X] = \sigma^2 = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x) dx$$

- Standard deviation

$$std[X] = \sigma = var[X]^{1/2}$$

- N-th moment

$$E[X^N] = \int_{-\infty}^{\infty} x^N f_X(x) dx$$

Random vectors

- An extension of the concept of a random variable
  - A random vector $\underline{X}$ is a function that assigns a vector of real numbers to each outcome ξ in sample space S
  - We generally denote a random vector by a column vector
- The notions of cdf and pdf are replaced by 'joint cdf' and 'joint pdf'
  - Given random vector $\underline{X} = [x_1, x_2 \dots x_N]^T$ we define the joint cdf as

$$F_{\underline{X}}(\underline{x}) = P_{\underline{X}}[\{X_1 \le x_1\} \cap \{X_2 \le x_2\} \dots \{X_N \le x_N\}]$$

  - and the joint pdf as

$$f_{\underline{X}}(\underline{x}) = \frac{\partial^N F_{\underline{X}}(\underline{x})}{\partial x_1 \partial x_2 \dots \partial x_N}$$

- The term <u>marginal pdf</u> is used to represent the pdf of a subset of all the random vector dimensions
  - A marginal pdf is obtained by integrating out variables that are of no interest
  - e.g., for a 2D random vector $\underline{X} = [x_1, x_2]^T$, the marginal pdf of x_1 is

$$f_{X_1}(x_1) = \int_{x_2=-\infty}^{x_2=+\infty} f_{X_1 X_2}(x_1 x_2) dx_2$$

Statistical characterization of random vectors

- A random vector is also fully characterized by its joint cdf or joint pdf
- Alternatively, we can (partially) describe a random vector with measures similar to those defined for scalar random variables
- Mean vector

$$E[X] = \underline{\mu} = \left[E[X_1], E[X_2] \dots E[X_N]\right]^T = [\mu_1, \mu_2, \dots \mu_N]^T$$

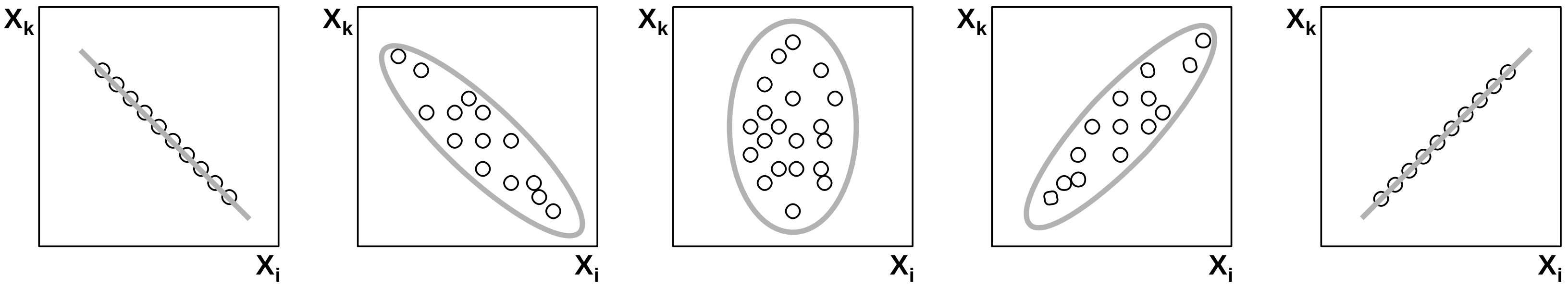
- Covariance matrix

$$cov[X] = \Sigma = E\left[\left(\underline{X} - \underline{\mu}\right)\left(\underline{X} - \underline{\mu}\right)^T\right] =$$

$$= \begin{bmatrix} E[(x_1 - \mu_1)^2] & \dots & E[(x_1 - \mu_1)(x_N - \mu_N)] \\ \vdots & \ddots & \vdots \\ E[(x_1 - \mu_1)(x_N - \mu_N)] & \dots & E[(x_N - \mu_N)^2] \end{bmatrix} =$$

$$= \begin{bmatrix} \sigma_1^2 & \dots & c_{1N} \\ \vdots & \ddots & \vdots \\ c_{1N} & \dots & \sigma_N^2 \end{bmatrix}$$

- The covariance matrix indicates the tendency of each pair of features (dimensions in a random vector) to vary together, i.e., to co-vary*
  - The covariance has several important properties
    - If x_i and x_k tend to increase together, then $c_{ik} > 0$
    - If x_i tends to decrease when x_k increases, then $c_{ik} < 0$
    - If x_i and x_k are uncorrelated, then $c_{ik} = 0$
    - $|c_{ik}| \leq \sigma_1 \sigma_k$, where σ_i is the standard deviation of x_i
    - $c_{ii} = \sigma_i^2 = var[x_i]$
  - The covariance terms can be expressed as $c_{ii} = \sigma_i^2$ and $c_{ik} = \rho_{ik}\sigma_i\sigma_k$
    - where $\rho_{ik}$is called the correlation coefficient



$C_{ik}=-\sigma_i\sigma_k$
$\rho_{ik}=-1$

$C_{ik}=-½\sigma_i\sigma_k$
$\rho_{ik}=-½$

$C_{ik}=0$
$\rho_{ik}=0$

$C_{ik}=+½\sigma_i\sigma_k$
$\rho_{ik}=+½$

$C_{ik}=\sigma_i\sigma_k$
$\rho_{ik}=+1$

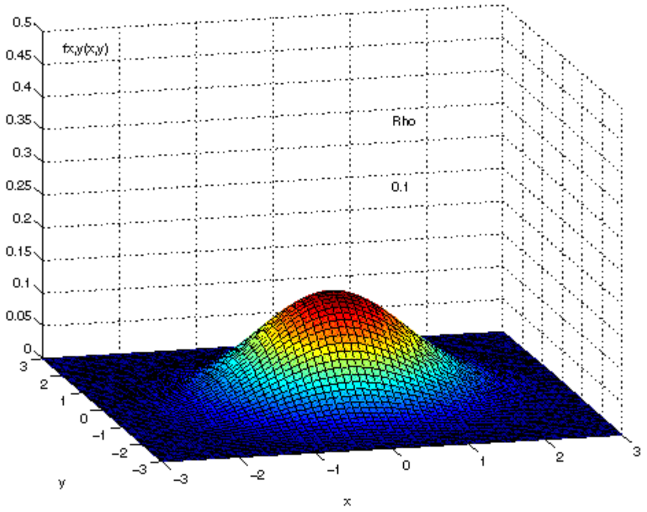
The Normal or Gaussian distribution

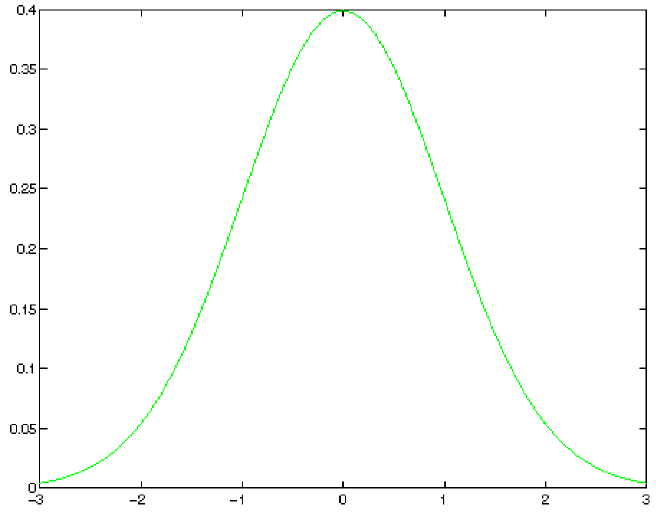
- The multivariate Normal distribution $N(\mu, \Sigma)$ is defined as

$$f_X(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

- For a single dimension, this expression is reduced to

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



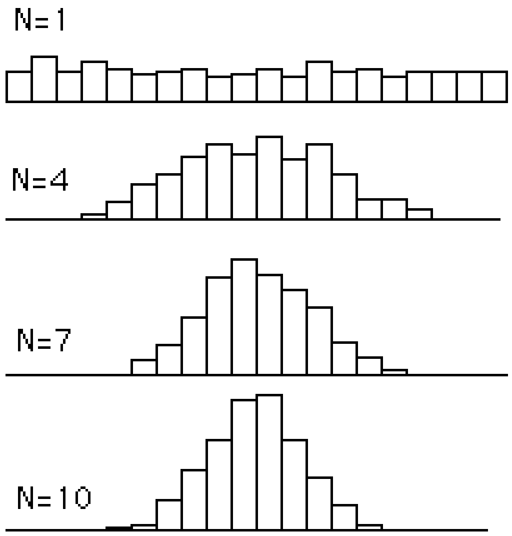


- Gaussian distributions are very popular since
  - Parameters (μ, Σ) uniquely characterize the normal distribution
  - If all variables x_i are uncorrelated $(E[x_i x_k] = E[x_i]E[x_k])$, then
    - Variables are also independent $(P[x_i x_k] = P[x_i]P[x_k])$, and
    - Σ is diagonal, with the individual variances in the main diagonal
  - Central Limit Theorem (next slide)
  - The marginal and conditional densities are also Gaussian
  - Any linear transformation of any N jointly Gaussian rv's results in N rv's that are also Gaussian
    - For $X = [X_1 X_2 \ldots X_N]^T$ jointly Gaussian, and $A_{N \times N}$ invertible, then $Y = AX$ is also jointly Gaussian

$$f_Y(y) = \frac{f_X(A^{-1}y)}{|A|}$$

Central Limit Theorem

- Given any distribution with a mean μ and variance σ^2, the sampling distribution of the mean approaches a normal distribution with mean μ and variance σ^2/N as the sample size N increases
  - No matter what the shape of the original distribution is, the sampling distribution of the mean approaches a normal distribution
  - N is the sample size used to compute the mean, not the overall number of samples in the data
- Example: 500 experiments are performed using a uniform distribution
  - $N = 1$
    - One sample is drawn from the distribution and its mean is recorded (500 times)
    - The histogram resembles a uniform distribution, as one would expect
  - $N = 4$
    - Four samples are drawn and the mean of the four samples is recorded (500 times)
    - The histogram starts to look more Gaussian
  - As N grows, the shape of the histograms resembles a Normal distribution more closely



Estimation theory

The estimation problem

- Suppose that a set of random variables $X = \{X_1, X_2 \dots X_N\}$ is iid (independent identically distributed) according to pdf $p(x|\Phi)$ but the value of Φ is unknown
- We seek to build an estimator of Φ, a real-valued function $\theta(X_1, X_2 \dots X_N)$ that specifies the value of Φ for each possible set of values of $X_1, X_2 \dots X_N$
- Three types of estimation procedures are commonly used
  - Minimum Mean Squared Error / Least Squares Error
  - Maximum Likelihood
  - Bayesian

Minimum Mean Squared Error / Least Squares Error

- Assume two random variables X and Y are iid according to $f_{xy}(x,y)$
- Suppose we do a series of experiments and observe the value of X
- We seek to find a transformation $\hat{Y} = g(X,\Phi)$ that allows us to predict the value of Y
  - This assumes that we know the general form of $g(\quad)$ but not the specific value of its parameters Φ
- The following quantity can measure the goodness of $\hat{Y} = g(X,\Phi)$

$$E\left(Y-\hat{Y}\right)^2 = E\left(Y-g(X,\Phi)\right)^2$$

  - This quantity is called the *mean squared error* (MSE)

- The process of finding parameter $\hat{\Phi}_{MMSE}$ that minimizes the MSE is known as the *minimum mean squared error (MMSE) estimator*

$$\hat{\Phi}_{MMSE} = \underset{\Phi}{\operatorname{argmin}}\left[E\left(Y-g(X,\Phi)\right)^2\right]$$

- In some cases, however, the joint pdf $f_{xy}(x,y)$ is unknown, so we must estimate Φ from a training set of samples (x,y)
- In this case, the following criterion can be used

$$\hat{\Phi}_{LSE} = \underset{\Phi}{\operatorname{argmin}} \sum\nolimits_{i=1}^{n} (y_i - g(x_i, \Phi))^2$$

- The process of finding parameter $\hat{\Phi}_{LSE}$ that minimizes this sum-squared-error (SSE) is called the *least squared error (LSE)* or *minimum squared error (MSE)* estimator

- We will now derive MMSE/LSE estimates for two classes of functions
  - Constant functions $G_c = \{g(x) = c; c \in \mathcal{R}\}$
  - Linear functions $G_l = \{g(x) = ax + b; a, b \in \mathcal{R}\}$

MMSE/LSE for constant functions

- When $\hat{Y} = g(x) = c$, the MSE becomes

$$E\left(Y - \hat{Y}\right)^2 = E(Y - c)^2$$

- To find the MMSE estimate of c, we take derivatives and equate to 0

$$c_{MMSE} = E(Y)$$

  - which indicates that the MMSE estimate is the expected value of Y
  - Likewise, it is trivial to show that the MSE is the variance of Y

- Following the same procedure, we find that the LSE estimate is

$$c_{LSE} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

  - which is the sample mean

MMSE/LSE for linear functions

- When $\hat{Y} = g(x) = ax + b$, the objective function becomes

$$e(a,b) = E\left(Y - \hat{Y}\right)^2 = E(Y - ax - b)^2$$

- To find the MMSE estimate for c, we take partial derivatives with respect to a and b and equate to 0

$$\frac{\partial e}{\partial a} = 0 \Rightarrow a = \frac{cov(X,Y)}{var(Y)} = \rho_{xy}\frac{\sigma_x}{\sigma_y}$$
$$\frac{\partial e}{\partial b} = 0 \Rightarrow b = E(Y) - \rho_{xy}\frac{\sigma_x}{\sigma_y}E(X)$$

- To find the LSE estimate, assume that we have n sample-vectors $(\boldsymbol{x}_i, y_i) = \left(x_i^1, x_i^2 \dots x_i^d, y_i\right)$
- The linear function can be represented as

$$\hat{Y} = XA$$

- Or in expanded form as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1^1 & & x_1^d \\ 1 & x_2^1 & & x_2^d \\ & & & \\ 1 & x_n^1 & & x_n^d \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{bmatrix}$$

  - where we have absorbed the intercept b by adding a constant dimension

- The SSE can then be represented as

$$e(A) = \left\|\hat{Y} - Y\right\|^2 = \sum\nolimits_{i=1}^{n} (A^T \boldsymbol{x_i} - y_i)^2$$

- A closed-form solution to this estimate can be obtained by taking the gradient of $e(A)$ and equating to 0

$$\nabla e(A) = \sum\nolimits_{i=1}^{n} 2(A^T \boldsymbol{x_i} - y_i)\boldsymbol{x_i} = 2X^T(XA - Y) = 0$$

- which yields the following form

$$A_{LSE} = (X^T X)^{-1} X^T Y$$

  - The expression $X^{\perp} = (X^T X)^{-1} X^T$is known as the pseudo-inverse of X

- When $X^T X$ is singular, the pseudo-inverse cannot be computed
- In this case, we use the alternative objective function

$$e(A) = \|XA - Y\|^2 + \alpha\|A\|^2$$

  - where α is known as a *regularization* parameter

- Following a similar procedure as before, we obtain the LSE estimate

$$A_{LSE} = (X^T X + \alpha I)^{-1} X^T Y$$

  - which is generally known as the *regularized LSE* or *ridge-regression* solution ex10p1.m

Find the LSE solution (1 dimensional model)

ex10p2.m

Find the LSE solution (3 dimensional model)

Maximum Likelihood Estimation (MLE)

- MLE is the most commonly used parametric estimation method
- Assume that a set of random samples $X = \{X_1, X_2 \dots X_n\}$ are *independently* drawn from pdf $p(x|\Phi)$
- Assume that we make a number of observations $x = (x_1, \dots x_n)$
- In MLE we seek to find the set of parameters Φ that maximize the observations
- Since $X = \{X_1, X_2 \dots X_n\}$ are independently drawn, the joint likelihood can be rewritten as

$$p_n(x|\Phi) = \prod_{k=1}^{n} p(x_k|\Phi)$$

- and the maximum likelihood estimate is

$$\Phi_{MLE} = \underset{\Phi}{\operatorname{argmax}}\, p_n(x|\Phi)$$

- Since the logarithm is a monotonically increasing function, we generally maximize the log-likelihood

$$\mathrm{l}(\Phi) = \log p_n(x|\Phi) = \sum_{k=1}^{n} \log p(x_k|\Phi)$$

MLE example

- Let’s look at the MLE for a univariate Gaussian

$$p(x|\Phi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

  - where in this case $\Phi = \{\mu, \sigma^2\}$

- The log likelihood is

$$\log p_n(x|\Phi) = \log \prod_{k=1}^{n} p(x_k|\Phi) =$$

$$\sum_{k=1}^{n} \log \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-(x_k-\mu)^2/(2\sigma^2)} \right] =$$

$$-\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k - \mu)^2$$

- Taking partial derivatives, setting to zero and solving for μ, σ^2yields

$$\mu_{MLE} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

$$\sigma_{MLE}^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \mu_{MLE})^2$$

- which shows that the MLEs for the mean and variance are the sample mean and the sample variance

Bayesian estimation

- Bayesian estimation follows a different philosophy from MLE
  - MLE assumes that the parameter Φ is unknown but fixed
  - Instead, BE assumes that the parameter Φ itself is a random variable with its own prior distribution $p(\Phi)$
- The most popular form of Bayesian estimation is the so-called *Maximum A Posteriori* (MAP) estimation
- Given observation sequence $\boldsymbol{x} = (x_1, \ldots x_n)$, the posterior distribution of Φ can be obtained using Bayes' rule as

$$p(\Phi|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\Phi)p(\Phi)}{p(\boldsymbol{x})} \propto p(\boldsymbol{x}|\Phi)p(\Phi)$$

- In MAP, we seek to find the parameter that maximizes $p(\Phi|\boldsymbol{x})$

$$\hat{\Phi}_{\text{MAP}} = \underset{\Phi}{\operatorname{argmax}}\, p(\Phi|\boldsymbol{x})$$

- The MAP estimator allows us to incorporate any prior knowledge we may have about parameter Φ by means of prior $p(\Phi)$
  - When the amount of data is limited, the MAP estimator relies more heavily on the prior $p(\Phi)$
  - As the amount of data increases, MAP begins to balance information in the prior and in the likelihood $p(\boldsymbol{x}|\Phi)$
  - For large enough n, MAP approaches the MLE solution
- If we set the prior $p(\Phi)$ to a constant value (also known as a *non-informative* prior), MAP estimation becomes equivalent to MLE